International Journal on Recent and Innovation Trends in Computing and Communication ISSN: 2321-8169
Volume: 3 Issue: 12 6572 - 6576

Benchmark Classification of Handwritten Dataset by New Operator

Rajiv Ranjan
M.Tech Scholar
SE, CS Branch, Suresh Gyan Vihar University, Jaipur, India
Rajivsh1992@gmail.com

Mohit Vats
Assistant Professor
CE&IT Branch, Suresh Gyan Vihar University, Jaipur, India
Mohit.vats@mygyanvihar.com

Sachin Jain
Assistant Professor
CE&IT Branch, Suresh Gyan Vihar University, Jaipur, India
Sachin.jain@mygyanvihar.com

***Abstract*—** In recent years, many new classifiers and feature extraction algorithms were proposed and tested on various OCR databases and these techniques were used in wide applications. Various systematic papers and inventions in OCR were reported in the literature. We can say that OCR is one of the most important and active research areas in the pattern recognition. Today, research OCR is dealing with diverse a character of complex problems. Important research in OCR includes the text degraded (heavy noise) and analysis/recognition of complex documents (including texts, images, graphs, tables and video documents).
In this proposed system we are suing a new operator Recognition of Devnagari handwritten Characters one of the biggest problem in present scenario. Devnagari characters are not recognized efficiently and truthfully by electronic device. Many researchers and algorithm have been proposed for recognizing of characters. For recognizing of characters, many processes have to be performed but no single technique or algorithm can perform that recognition and give more accurate result.
objective of this dissertation work is to propose a new operator, the name of this operator is Kirsch Operator and algorithm for getting accurate result.

***Keywords*-**OCR, Kirsch Operator,

*****

I. INTRODUCTION

In recent years, many new classifiers and feature extraction algorithms were proposed and tested on various OCR databases and these techniques were used in wide applications. Various systematic papers and inventions in OCR were reported in the literature. We can say that OCR is one of the most important and active research areas in the pattern recognition. Today, research OCR is dealing with diverse a character of complex problems. Important research in OCR includes the text degraded (heavy noise) and analysis/recognition of complex documents (including texts, images, graphs, tables and video documents).

Optical Character Recognition (OCR), branch of image processing, pattern recognition, and computer vision. More than five decades OCR has been broadly researched. With the arrival of digital computers many researchers and engineers get involved in this remarkable topic. It is not only a recently mounting topic because many potential applications, as the bank check the postal sorting mail procedure, automatic reading of tax forms and miscellaneous manuscript materials Printed, but it is also a reference for the testing and verification of algorithms and new theories of recognition motif. Handwritten character recognition, (as there are varieties of writing styles according to the age of the applicant, sex, education, ethnicity, etc., as well as the mood of the writer when writing), is an area of quite difficult research in OCR. These challenges attract the researcher in this field.

Recognition of Devnagari handwritten Characters one of the biggest problem in present scenario. Devnagari characters are not recognized efficiently and truthfully by electronic device. Many researchers and algorithm have been proposed for recognizing of characters. For recognizing of characters, many processes have to be performed but no single technique or algorithm can perform that recognition and give more accurate result. In the today scenario, it is more important recognizing of Devnagari characters because in India Hindi is mother language. This system helps human being to solve their more complex problem in very easy way. In active area of research hand written characters is a problem of the recognition. Because it is very important requirement of office automation. It is provide to effective and practical reorganization of characters. At the time of writing of a person depends on their moods and writing styles its does not lend .It help in recognition process and all structure, statistical and topological information about the character in all, the sort has been observed in the recognition process. In the Hand written Hindi characters limited variation and shape and size are consider and main attentive focus on the recognition.

35 years passed away, that researchers had been working on Hand written recognition .From last years, characters of compares those were there in the research on the hand written recognition are gaining continually. Nowadays, public has become attentive towards the handwritten recognition technology. aim of the creating handwritten recognition system with the rating of 100% is still not achieved as humans beings is not possible to recognized every test of any writer without any confusion most of the people cannot read their own hand writing in the very effectively . It is the responsibility of the writer to write the text in the readable format.

II. LITERATURE REVIEW

Development of the system of optical character recognition (OCR) for handwritten character recognition without constraint in offline mode is active, yet challenging field of research [1]. Variations in handwriting pose a major challenge in the development of accurate recognition system. The reason is that handwriting changes induce virtually unmanageable variations which make the definition of extremely difficult feature. Thus, discover a precise and efficient feature extraction method has become a daunting task. However, to deal with it, there have been several efforts to define and extract features that may have reduced the effects of variations in handwriting. In this piece of writing, we present our experiences with the domain feature transformed according to definition and extraction techniques in the development of a system for the recognition of the figures without constraint of handwritten epithets. that the performance of structural features ranging from 40 to 97%. Only in one case [2], it is reported 99.5%, while the accuracy stated statistical characteristics [3] is in the order of 40 to 92%, which is a little less. The precision of the type of function space, as the binary image [8-9] and box [6] also reported precision series between 95 to 96%. Table-II shows some results from field transformed using wavelets [10] and degraded features [11, 12]. The result shows that the result of the recognition in this case is 94.25 at 99.56, which is slightly higher than the approach of the spatial domain. As mentioned previously, we thoroughly explored performance of the features of this area on a reference dataset significantly larger, more realistic and more uniform. Aware of the importance of such a set of data, Centre of excellence for the analysis of documents and recognition Cedar [39] has launched an ambitious project ("creation of data resources and the design of a test bench for evaluation for recognition of characters of epithets" (from the collection, the compilation of a test data set field.) These data are available on CDROM. It contains the handwritten, postal codes, characters and alphabetic characters. It is not available in the public domain. It can be purchased. In addition, the attributes of data that were considered during the collection of data are not available to the public. In 2007, Pal et al. [Ref] developed datasets for handwritten figure for six languages. They collected: 22546, 14650, 2220, 5638, 4820 and 2690 figures sample scripts epithets, Bengali, Telugu, Oriya, Kannada, and Tamil respectively. The same year Chaudhary et al. [4] had a developed characters handwritten Bangle online samples. The data sets include 7000 sets learning and test 1348 together. Blais et al. [5] in the year 2009 reported the creation of a dataset for the figures and characters isolated. A year later, in 2010, Genevieve et al. [6] had developed datasets for Kanata and Tamil languages. This set was created by 600 subjects and it contains 100000 words. In 2011, a handwritten document Kannada KHTD data group [7] was collected by Alaei et al. [29]. Their Dataset contains 204 manuscripts, the lines of text 4298 and 26115 words of four different categories written by 51 native speakers of Kannada. The total character of lines of text and words in data sets is 4298 and 26115 respectively. The same year a dataset that contains 26 720 words handwritten legal amount written in Hindi and Marathi (Devnagari script) was created by Jerome et al. [31]. Recently in 2012 CMATERdb1 [40] had also created a data sets for the manuscript text of Bangle and Bangle text mixed of English words.

III. OBJECTIVE OF OUR WORK

The dataset includes isolated manuscript figures, characters, handwritten constrained and unconstrained. The dataset is organized in a database as well as information and related attributes in writing writers intending to facilitate the statistical analysis of Scripture against the demographic variables. The novelty of this data set is that handwritten speech samples have been reproduced from a pangram: a sentence in which each letter of the alphabet appears at least once. The pangram is designed specifically for this research captures the possible variations in handwriting epithets. Samples were taken of different persons in different age groups, ethnic origin, and level of training and regional and linguistic groups. The dataset contains 49000 (30000 constrained and unconstrained 19000) handwritten character, isolated characters handwritten 82609 and 1700 pangram forced without constraint and 1700 text. Below, the dataset will be referred to as CPAR-2012 data set.

This set of data provided an environment of experimentation estimate: (a) recognition accuracy: time of recognition (b) and (c) training size effect. The estimation of (a) to (c) is as realistic as possible. Estimated time includes the time required by recognizing a character of preprocessing in the standings on Intel ® core ™ 2 Duo CPU 2.00 GHZ, 64-bit, OS, x-64 base processor with 4.00 GB of RAM and MTALAB R2015a. To the best of our information no work epithets character mentioned the statistics on the throughput of the system. These tables shows also that almost all Hindi OCR reported techniques have been tested and experimented with synthetic (hand created or simulated) [2-9, 12] of data sets.

The database contains 150 handwritten pages, among the 100 pages are purely written in Bangla and rests of the 50 pages are written in Bangla text mixed with English words. From the data, mentioned above, we observe that very little attention has been paid to the collection of test data by researchers at the beginning of recognition of characters Devnagari. In our evaluation lack of data sets standard is a major reason for the slow development in the context of epithets document recognition. Another equally important reason is the complex structure of the characters epithets that it is difficult to describe or represent in computer form.

IV. PROPOSED WORK

Linear discriminate analysis can be used for the classification of the model. LDA handles the case where the class frequencies are uneven and their performance has been tested on randomly generated test data sets. This classifier maximizes the ratio of the variance between the classes to the intraclass variance in a particular data set, thus guaranteeing a maximum separation. The classifier uses two approach for classification-class - dependent transformation and class independent transformation. Transformation dependent class is to maximize the relationship between the

variance of class indoors class gap. The main objective is to maximize this ratio so that sufficient classs separability can be obtained. Independent class Transformation approach is to maximize the relationship between the overall variance indoors class gap. This approach uses only a single criterion for optimization to transform data sets and therefore all data regardless of their class identity points are transformed using this transformation. In this type of LDA, each class is considered a separate category against all other classes

1 ALGORITHM

Step1: Load CPAR datasets.
Step2: Resize each image of size 32X32.
Step3: Apply horizontal, vertical, left-diagonal and right-diagonal kirsch operator to extract four local feature vectors and one global feature through resizing the image directly.
Step4: Apply the 'db1' wavelet transformation of the filtered images.
Step5: Combine both feature vectors (local and global) to form a final feature vector of length 80.
Step 6: Classify the feature vector by using two statistical classifier LDA and KNN.
Step6: Find Recognition Results.

V. IMPLEMENTATION AND RESULTS

Several experiments have been conducted to assess the quality of the CPAR dataset. Initially, we experimented with handwritten characters. In these features experiences from simple profile and served to the operator of Kirsch.
Sridhar and Anita [45] profile features used for the recognition of characters in 1984. They extracted characteristics of right and left profiles. Hutted et al. [46] use left, right, top and bottom profile for handwritten mineralogical. These profiles to the character 3 are shown in figure 21. Detailed algorithm is shown in figure 21.
.

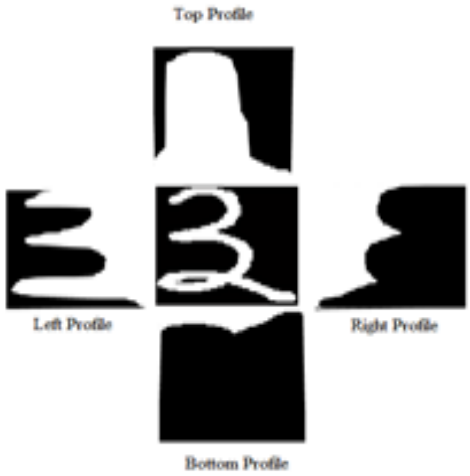

Figure 1.1:- Simple profile Recognition Algorithm

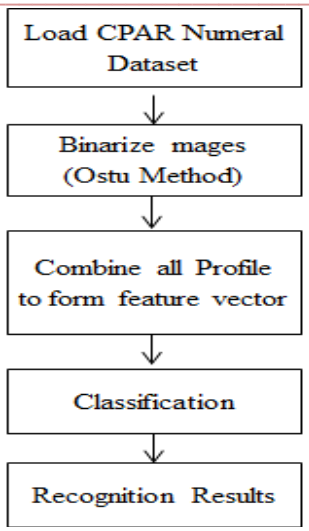

Figure 1.2:- Simple profile

A gradient operator is used to find the degraded functionality. The gradient used in character recognition operator includes operator Robert [47], the Sobel operator [48] and operator of Kirsch [49]. However, among these operators, the operator of Kirsch has been known to detect the four directional edges more precisely that the other operator because Kirsch operator considers all eight directions. The figure shows eight masks used in the present. Kirsch operator uses eight masks to calculate components degraded in eight directions. (24) the figure shows four masks and four other masks are symmetrical with that. In this article, vectors of the directional characteristic for horizontal (H), vertical (V), diagonal right (R) and the left diagonal (L) indications are calculated as following:

$G(i, j)_H = \max(|5S_0 - 3T_0|,|5S_4 - 3T_4|)$,
$G(i,j)_V = \max(|5S_2 - 3T_2|, |5S_6 - 3T_6|)$,
$G(i,j)_R = \max(|5S_1 - 3T_1|, |5S_5 - 3T_5|)$,
$G(i, j)_L = \max(|5S_3 - 3T_3|, |5S_7 - 3T_7|)$, - (2)

Where $S_k = A_k + A_{k+1} + A_{k+2}$ (3)
$T_k = A_{k+3} + A_{k+4} + A_{k+5} + A_{k+6} + A_{k+7}$ (4)
Eight neighboring pixels are shown in Figure 22

A_0	A_1	A_2
A_7	(i, j)	A_3
A_6	A_5	A_4

One of the main reasons to use the operator of kirsch, is that it gives directly the strength of all eight directions. However given that masks are not orthogonal i.e. a specific direction mask does not zero force degraded to the edge in a direction perpendicular, the directional elements are not well separated [. The local characteristics were extracted by applying three levels Daubechies (db-1) Wavelet on four images that were obtained by applying the operator of Kirsch. The transformed Wavelet compress the original image standard 32 X 32 in size from the image of 4 x 4, total

it creates 4 x 4 x 4 = 64 local device. The global characteristics have been obtained by applying the same transformation on the standardized original image (32 x 32) and it creates 4 x 4 overall functionality. Vector's last feature was formed by combining 4 x 4 x 4 = 64 local and global 4 x 4 = 16 function of the total length of 5 x 4 x 4 = 80, as shown in figure (23). A characteristic vector of length 80 was formed by combining the capabilities of global 16and local 64. Figure 25 shows the detailed algorithm.

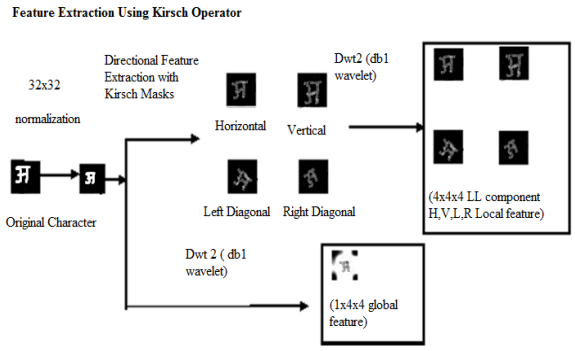

Figure 1.3:- Feature extraction using Kirsch Operator

We have developed a set of standard data for recognition of handwritten characters without constraint (CPAR Datasets). Which contains handwritten isolated figures 49000, 30000 figures were used for training purposes and 19000 figures were used for test and it includes also 82609 handwritten characters whose 50379 handwritten characters may be used for training purposes and 32230 can be used for testing purposes..Detailed structure

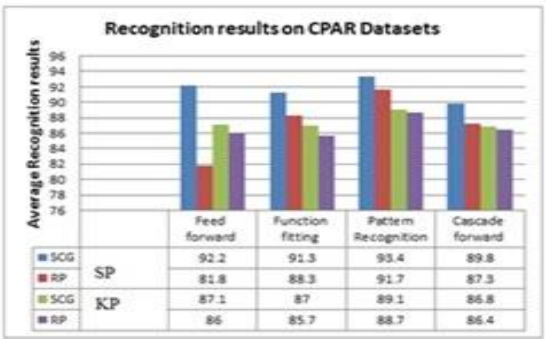

Figure 1.4:- Recognition results of CPAR datasets RP: Resilient Back propagation,

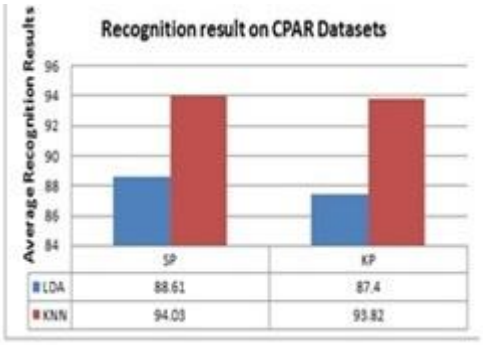

Figure 1.5:- SCG: Scaled Conjugate Gradient back propagation SP: Simple profile , KP: Kirsch

It is clear that network model classifier neural recognition gives better results compared to other neural networks. He also observed that when this network formed with gradient conjugate scaling algorithm gives better results. Since the algorithm of gradient combined with scaling (SCG) was designed to avoid tedious it is also faster than the other classifier neural network. In the case of the waterfall before the neural network during training with elastic spread back observed need more calculation and more time. Elastic back multiplication algorithm must more calculation to update weight is also slower than the recognition neural network model. Compared to the statistical classifier KNN gives the best results of all the other above mentioned classifier. Experiment with Devnagari Characters.

VI. CONCLUSION

To develop a high performance integrated research environment a large scale datasets contributed by thousands or even more may be needed. The Multi-type offline Devnagari character datasets has an important feature in these datasets that are not available in other Devnagari datasets. It is the first datasets that covers isolated handwritten numerals, isolated handwritten characters, constrained handwritten text, and unconstrained handwritten text written by more than 1700 persons from different backgrounds. Writers' information was also stored in the datasets for retrieving datasets by different filtering criteria. The collection of handwriting sample never completes, since there is large variation in handwriting at different level of age as well as at different generation. For uniform progress in this area more and more challenging data will become necessary. We are still working on this to build handwriting sentence datasets which are in the progress and by the end of this year we may complete this. Finally we can state that data collection is an endless journey.

VII. FUTURE WORKS

We can use this algorithm in the Devnagari text and we can use this algorithm in unconstrained characters. This algorithm can further be used in broken characters. In online character reorganization this algorithm can also be used.

VIII. REFERENCES

[1] Jawad AlKhateeb, Jinchang Ren. Jianmin Jiang, Husni Al. Muhtaseb, ―Offline handwritten Arabic cursive text recognition using Hidden Markov Models and Re-ranking‖ in Pattern Recognition vol.32, pp.1081-1088, 2011.

[2] Ved Prakash Agnihotri, ―Offline Handwritten Devnagari Script Recognition‖ in MEC, pp. 37-42, 2012.

[3] U. Pal and B. B. Chaudhuri, ―Indian script character recognition: A survey,‖ Pattern Recognition., vol. 37, pp. 1887–1899, 2004.

[4] R. Jayadevan, Satish R. Kolhe, Pradeep M. Patil and Umapada Pal, ―Offline Recognition of Devanagari Script: A Survey‖ in IEEE Transaction on Systems, Man and Cybernetics –Part C. Applications and Review, vol.41, pp-782-796, 2011.

[5] Bikash Shaw, Swapan Kr. Parui and Malayappan Shridhar, ―Offline Handwritten Devanagari Word Recognition: A Segmentation Based Approach‖, 19th

International Conference on Pattern Recognition (ICPR'08), December, 8-11, 2008, Tampa, Florida, USA.

[6] Naveen Shankaran, Aman Neelappa and C.V. Jawahar, ―Devanagari Text Recognition: A Transcription based Formulation‖ in ICDAR, pp. 678-68, 2013.

[7] Vedgupt Saraf, ―Offline Handwritten Character Recognition of Devanagari script uses Genetic Algorithm for Improve efficiency‖ in ICCSE, pp.161-164, 2013.

[8] A. Bharat and Sriganesh Madhavnath, ―HMM – Based Lexicon Driven and Lexicon-Free word Recognition for Online Handwritten Indic Scripts‖ in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol-34, pp.670-682, 2012.

[9] Sandhya Arora and Debotosh Bhattacharjee, ―Multiple classifier combination for Offline Handwritten Devanagari Character Recognition‖.

[10] Umapada Pal, T. Wakabayashi, F. Kimura, ―Comparative study of Devanagari Handwritten Character Recognition using Different Features and Classifiers‖ in 10th ICDAR, IEEE, pp.1111-1115, 2009.